FlowKV Isolated Cache Management and Inference Throughput in Llama-3-8B Conversational Tasks

Assignee Research

June 7, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: What is the impact of FlowKV's isolated KV cache management on the inference throughput of Llama-3-8B compared to standard KV cache eviction methods during multi-turn conversational tasks. 12 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: FlowKV: Enhancing Multi-Turn Conversational Coherence in LLMs via Isolated Key-Value Cache Management. Research question: What is the impact of FlowKV's isolated KV cache management on the inference throughput of Llama-3-8B compared to standard KV cache eviction methods during multi-turn conversational tasks?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

11 papers retrieved. 12 claims extracted; 0 independently verified. Quality review score: 4.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
FlowKV achieves an Instruction Following Rate (IFR) of 61.93% on Turn 2 for the LLaMA model using the SKV strategy, comp	×	0.05
FlowKV achieves an Instruction Following Rate (IFR) of 54.95% on Turn 3 for the LLaMA model using the SKV strategy, repr	×	0.05
On the LLaMA model with the CKV strategy, FlowKV improves Turn 2 IFR by 40.27% over the baseline (52.83% vs 12.56%).	×	0.03
FlowKV achieves an average performance improvement of over 20% in subsequent conversation turns compared to the baseline	×	0.04
During the initial turn of conversation, FlowKV's core isolation mechanism is not engaged due to the absence of prior co	×	0.11
In a 3-turn dialogue, Turn 1 Response attention is heavily focused on Turn 1 Query (T1Q) and the local window.	×	0.04
In a 3-turn dialogue, Turn 2 Response attention expands to include Turn 1 Query (T1Q), Turn 1 Response (T1R), and the lo	×	0.04
Queries in Turns 2 and 3 exhibit increased attention to previous queries and the initial system prompt.	×	0.02
FlowKV achieves a Turn 2 IFR of 56.72% on the Qwen model using the SKV strategy, an improvement of +39.39% over the base	×	0.02
SnapKV and ExpectedAttention exhibit minimal performance degradation when used with FlowKV compared to Full KV.	×	0.05
The Full KV Baseline achieves an IFR of 75.4% on the PrefEval benchmark.	×	0.04
FlowKV achieves an IFR of 58.7% on the PrefEval benchmark.	×	0.02

References

- http://arxiv.org/abs/2605.07234v1
- http://arxiv.org/abs/2605.08840v1

- http://arxiv.org/abs/2505.15347v2